Rethinking Evaluation in Retrieval-Augmented Personalized Dialogue: A Cognitive and Linguistic Perspective

Tianyi Zhang, David Traum
Institute for Creative Technologies, University of Southern California
12015 Waterfront Drive, Los Angeles, CA 90094-2536, USA
tzhang62@usc.edu, traum@ict.usc.edu

Abstract

In cognitive science and linguistic theory, dialogue is not seen as a chain of independent utterances but rather as a joint activity sustained by coherence, consistency, and shared understanding. However, many systems for open-domain and personalized dialogue use surface-level similarity metrics (e.g., BLEU, ROUGE, F1) as one of their main reporting measures, which fail to capture these deeper aspects of conversational quality. We re-examine a notable retrieval-augmented framework for personalized dialogue, LAPDOG, as a case study for evaluation methodology. Using both human and LLM-based judges, we identify limitations in current evaluation practices, including corrupted dialogue histories, contradictions between retrieved stories and persona, and incoherent response generation. Our results show that human and LLM judgments align closely but diverge from lexical similarity metrics, underscoring the need for cognitively grounded evaluation methods. Broadly, this work charts a path toward more reliable assessment frameworks for retrieval-augmented dialogue systems that better reflect the principles of natural human communication.

Keywords: Personalized Dialogue Evaluation, Retrieval-Augmented Generation (RAG), Discourse Coherence and Common Ground

1. Introduction

Dialogue is not a sequence of isolated utterances; it is a collaborative process in which speakers build and maintain common ground (Clark, 1996; Clark and Brennan, 1991), manage discourse structure across turns (Grosz and Sidner, 1986), and coordinate meaning through linguistic and conceptual alignment (Pickering and Garrod, 2004; Poesio and Rieser, 2001; Mao et al., 2025). These requirements highlight why dialogue is a challenging task for natural language processing (NLP). Systems must preserve coherence over long and sometimes discontinuous histories, remain consistent with persona information, and avoid contradictions that would disrupt common ground (Zhu et al., 2023). One promising strategy for mitigating these issues is to give dialogue models access to external or supplementary information beyond the immediate conversation. Retrieval-augmented generation (RAG) implements this idea by combining neural retrieval with response generation, allowing the model to dynamically access relevant background or persona-related content.

We examine one recent approach to personalized dialogue generation, namely LAPDOG (Learning Retrieval Augmentation for Personalized Dialogue Generation) (Huang et al., 2023), which augments persona profiles by retrieving additional external stories. We critically analyze and re-evaluate LAPDOG, considering issues with the history used, the evaluation metrics, and contradictions and coherence with both the dialogue history and external materials. We compare standard similarity metrics with judgments from both human annotators and LLM evaluators to assess whether current evaluation practices capture the cognitive and discourse-level qualities that make dialogue coherent and meaningful. Our results show that while similarity metrics report gains, human and LLM evaluators reveal inconsistencies—particularly in coherence and persona consistency—highlighting a disconnect between surface overlap and communicative quality.

Our main contributions are: **(1) Evaluation methodology:** We introduce a systematic framework that combines human and LLM judges, and analysis methods grounded in cognitive and linguistic theories of dialogue. This framework highlights key aspects of conversational quality—such as coherence, persona consistency, and engagement—that reflect how humans maintain common ground in interaction. **(2) Empirical findings:** We show that human and LLM judgments align closely but diverge sharply from surface metrics, revealing that improvements measured by BLEU, ROUGE, and F1 often fail to reflect gains in communicative or cognitive quality. This divergence highlights a deeper evaluative gap between linguistic form and interactive function, pointing to the need for metrics grounded in discourse and cognitive theory. **(3) Future directions:** We outline paths toward cognitively grounded evaluation, including taxonomies of persona–story relations and multi-objective training objectives that more directly model coherence, engagement, and the maintenance of common ground.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11404–11416
11–16 May 2026.

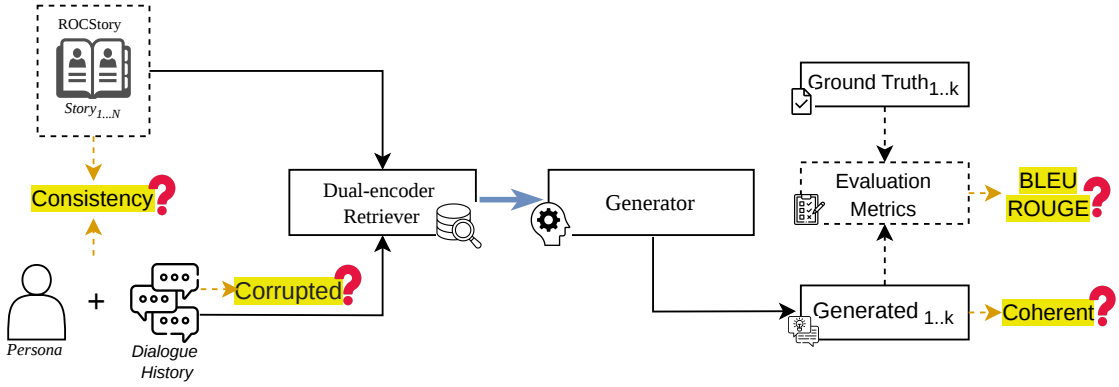

Figure 1: Overview of the LAPDOG retrieval-augmented personalized dialogue framework. The model retrieves external stories (e.g., from ROCStory) based on persona and dialogue history using a dual-encoder retriever, integrates them to a generator, and evaluates responses with metrics such as BLEU and ROUGE. The question marks indicate issues identified in our analysis—namely, use of similarity metrics, persona–story contradiction, incoherent generation, and dialogue corruption.

2. Related Work

Psycholinguistic theories of dialogue highlight how interlocutors update shared mental models and repair misunderstandings to preserve coherence (Clark, 1996; Brennan and Clark, 1996), while discourse theories emphasize the role of long-range dependencies and topic continuity in guiding interpretation (Grosz et al., 1995; Hobbs, 1979). Cognitive models of alignment further suggest that successful dialogue depends on mutual adaptation at both linguistic and conceptual levels (Garrod and Pickering, 2009; Pickering and Garrod, 2013). Together, these perspectives show that dialogue is not only about producing locally appropriate sentences but also about maintaining consistency, coherence, and engagement across extended interactions.

2.1. RAG for Personalized Dialogue

Retrieval-Augmented Generation (RAG) has emerged as an effective framework for enhancing personalized dialogue by retrieving external content—such as documents, knowledge sentences, or narratives—and conditioning the generator on that content. These methods aim to go beyond predefined persona profiles by supplementing limited persona and dialogue history with richer context. By grounding responses in retrieved information, RAG can help maintain coherence over long dialogues, reinforce persona consistency, and reduce contradictions that stem from limited contextual awareness.

Several recent RAG-based approaches, including UniMS-RAG (Li et al., 2024) and SAFARI (Wang et al., 2023), support multi-source integration by retrieving from persona memories, user profiles, and factual documents. PK-ICR (Oh et al., 2023) further improves retrieval by jointly selecting persona and knowledge pairs. However, these models generally assume access to sufficiently detailed persona stores and do not specifically target the problem of persona sparsity.

2.2. LAPDOG

LAPDOG (Learning Retrieval Augmentation for Personalized Dialogue Generation) (Huang et al., 2023) augments persona profiles by retrieving additional external stories. Retrieval candidates are selected using a dot-product similarity score between stories in a database and a persona-based dialogue query, and the retriever and generator are trained jointly with optimization based on similarity metrics. This approach improves performance on BLEU, ROUGE, and F1 compared to baseline models. However, these metrics emphasize surface-level lexical overlap and fail to capture deeper qualities of conversational competence such as coherence, engagement, and persona consistency (Liu et al., 2016a).

LAPDOG (Huang et al., 2023) is distinctive in its use of narrative story data to enrich sparse personas through a fully end-to-end RAG framework. Majumder et al. (2021) also retrieve stories from the ROCstory dataset (Mostafazadeh et al., 2016), though their approach uses fixed similarity retrieval rather than a jointly optimized retriever–generator pipeline. More recent work has explored adaptive retrieval and discourse grounding for improved contextual coherence—e.g., Ye et al. (2023) propose a retrieval–generation synergy model that dynamically updates persona representations during conversation, and Zhan et al. (2024) extend RAG with discourse-level contextualization to enhance turn-level consistency. Meanwhile, neuro-symbolic approaches have begun to bridge symbolic rea-

soning and neural retrieval for personalization; for instance, Zhu et al. (2024b) integrate structured user representations with neural affective models to capture individual variation in sentiment and intent, while other recent systems adopt hierarchical or personality-guided architectures to ensure consistency and emotion alignment in the dialogue generation (Zhu et al., 2024a; Xie and Mao, 2025).

LAPDOG's design illustrates the potential of retrieval augmentation. Its reliance on surface-level similarity metrics and certain structural limitations—such as corrupted histories, contradictions, and irrelevant retrievals—raise concerns that resonate with insights from cognitive science and linguistic theory. corrupted or fragmented dialogue histories prevent the model from maintaining the discourse structure that supports coherence across turns (Grosz et al., 1995). Similarly, when retrieved stories conflict with persona facts, the system breaks the common ground shared between speakers, which is essential for successful communication (Clark, 1996). Moreover, evaluation based only on lexical overlap neglects the cognitive and pragmatic dimensions of dialogue quality, such as engagement, consistency, and reasoning over shared knowledge.

2.3. Personalized Dialogue Evaluation

Many recent works in personalized or open-domain dialogue generation continue to report similarity metrics, such as BLEU, ROUGE, and F1 as their primary evaluation measures. For instance, Tang et al. (2023) evaluate contrastive latent-variable models for personalization largely through BLEU and ROUGE, while Lu et al. (2023) adopt similar metrics for assessing multi-attribute control in personalized dialogue. Even broader dialogue modeling efforts, such as Cheng et al. (2024) on in-dialogue learning, rely on these similarity-based scores as central indicators of quality. Although convenient, such metrics emphasize lexical overlap and provide little insight into discourse-level properties of conversation. From a cognitive science and linguistic theory perspective, this creates a misalignment: measures of surface similarity cannot capture coherence, consistency, or the maintenance of common ground that underlie natural dialogue (Grosz et al., 1995; Clark, 1996).

Several studies have shown that n-gram overlap metrics (e.g., BLEU, ROUGE, METEOR) poorly reflect human judgments in dialogue evaluation. Liu et al.(2016) find these metrics fail to distinguish high-quality responses from random or baseline outputs (Liu et al., 2016a), while Lowe et al.(2017) and Sharma et al.(2017) report similarly weak correlations in open-domain settings (Lowe et al., 2017; Sharma et al., 2017). These results highlight the need for evaluation methods that better capture coherence, persona consistency, and contextual relevance in personalized dialogue.

Recent studies have shown that large language models (LLMs) can align well with human judgments across a variety of tasks (Chiang and yi Lee, 2023; Wang et al., 2025). However, this alignment is not universal. For instance, Siro et al. (2024) found notable discrepancies between human and LLM evaluators in how user feedback was weighted during utterance evaluation. Similarly, Reiss (2023) cautioned that ChatGPT's utility in text classification is contingent on validation, due to task-specific inconsistencies. These findings stress the importance of validating LLM-based evaluation on a case-by-case basis. (Bavaresco et al., 2025) survey many tasks and models and conclude that LLMs should be carefully validated against human judgments before being used as evaluators. To our knowledge, no prior work has established whether LLMs exhibit strong agreement with humans in both Likert ratings and relative rankings when evaluating persona-grounded dialogue responses.

3. Critical Analysis of Retrieval-Augmented Personalized Dialogue

To better understand the challenges of retrieval-augmented approaches for personalized dialogue, we use LAPDOG as a representative case study. While LAPDOG demonstrates the promise of enriching persona profiles through external story retrieval, a closer look reveals several limitations, as illustrated in the system overview in Figure 1. In the subsections that follow, we analyze four key aspects: the reliance on supervised similarity metrics, incoherent response generation, contradictions between persona information and retrieved stories, and discontinuities caused by corrupted dialogue histories.

3.1. Use of Supervised Similarity Metrics

LAPDOG was evaluated using F1, BLEU, and ROUGE-L to measure similarity to a reference response given the same context and persona. These metrics were also used during training and in the retrieval process for the generator. This framework implicitly assumes that the reference response represents the best possible outcome. While it may be reasonable to assume that a crowd-sourced human response is superior to most randomly generated machine outputs, it is not clear that lexical or even semantic similarity constitutes a meaningful dialogue evaluation function (Liu et al., 2016b). Distinct responses can be equally or even more appropriate, while highly similar responses may be problematic due to a variety of pragmatic or con-

textual factors. Table 1 illustrates examples of a problematic response with high similarity metrics and an appropriate response with low similarity metrics.

Case	Example
High Similarity Metrics	*Contradictory Response* **Reference:** Yes, I love going to Disney World every year. **Model:** No, I love going to Disney World every year. **BLEU / ROUGE-L / F1:** 89.32 / 88.89 / 88.89
Low Similarity Metrics	*Appropriate Response* **Reference:** I enjoy visiting Disney World annually. **Model:** Yeah, we go there every year — it's my favorite trip. **BLEU / ROUGE-L / F1:** 3.09 / 0.00 / 0.00

Table 1: Illustrative limitations of surface-level similarity metrics. The first case shows high lexical overlap despite semantic contradiction. The second shows low lexical overlap despite pragmatic appropriateness.

R Persona	I am a recluse. I am a famous twitch streamer. I am working on a building a robot to be my body double. I helped design the game starcraft. I've an allergy to water.
Dialogue	**R:** Hi. I'm Billy. I work as an electrician in Texas. **Q:** Really? How? I do a lot of work on my mom and dads farm.
LAPDOG (as R)	I'm a recluse and I build robots. I'm trying to be my body double.

Table 2: Abrupt introduction of persona content leading to incoherent response.

3.2. Incoherent Response Generation

Like other persona-driven dialogue systems, LAPDOG's responses sometimes introduce information from the persona abruptly, resulting in a lack of contextual coherence with the preceding dialogue. For instance, as shown in Table 2, the LAPDOG system outputs the statement "I'm a recluse and I build robots. I'm trying to be my body double." after a discussion of being an electrician in Texas. While this line reflects a relevant persona trait, its sudden appearance feels disconnected from the conversation's flow and is phrased awkwardly. This kind of jarring insertion can break the naturalness of the dialogue and leave hearers wondering how reclusiveness relates to electrical work. Such examples underscore the need for better integration strategies that can ground persona-related content in a conversational context more fluidly. Better would be either showing relevance to the previous discussion (Sperber and Wilson) or explicitly signaling a proactive change in topic.

Persona	I love to sing songs from the movie *Frozen*. **I'm in the third grade.** I play with my friends on the playground. I love to go to Disney World every year. Mickey Mouse is my favorite character.
Retrieved Story	I have always wanted to play a character at Disney World. **Disney hired me as a customer service rep.** I moved to Orlando and applied for a job. I worked very hard to achieve my goal. The other day I got a promotion to play Mickey Mouse.

Table 3: Inconsistency between persona and retrieved story content. Contradictory story lines are bolded.

3.3. Persona and Retrieved Story Contradictions

Another significant issue involves the contradictions that arise between the external stories recovered and the established personas. For example, in Table 3, the retrieved story does in fact share multiple features in common with the persona, such as Disney World and Mickey Mouse. But the persona indicates that the speaker is in the third grade, while the retrieved story describes an adult worker. This mismatch could lead to responses that break character, ultimately reducing the believability and personalization of the system. Such inconsistencies highlight the importance of retrieval filtering, relevance scoring, or more sophisticated integration mechanisms that can enforce alignment between persona and retrieved content.

3.4. Discontinuous Dialogue History

We identified that the original LAPDOG implementation removes and reassigns some utterances from the dialogue history during both training and inference. For example, Table 4 illustrates a problematic training instance where LAPDOG constructs a new conversation using only utterances from line 2, 4, 6 and 7 of the original dialogue. This corruption introduces speaker inconsistencies and contextual

Line	Full Dialogue Context
1	R: Cars are my thing. Vintage cars. I love working on them. Wrestling? Do you enjoy it?
2	Q: Yes, I love the crowds, getting to know people.
3	R: I didn't think about the crowd aspect of wrestling. I do not like crowds.
4	Q: Understood. When i need to be alone, I work out a few times each week.
5	R: I agree. Working out is a great way to burn off steam. Do you like country music ?
6	Q: A little bit. I can get into taylor swift.
7	R: Lol. She's more pop now. Her old stuff was country. I like classic country.
Line	**Corrupted Input Used by LAPDOG**
2	Q: Yes, I love the crowds, getting to know people.
4	R: Understood. When I need to be alone, I work out a few times each week .
6	Q: A little bit. I can get into Taylor Swift .
7	R: Lol. She's more pop now. Her old stuff was country. I like classic country .

Table 4: Example original and corrupted dialogue.

disconnects. In the original dialogue, line 4 the utterance " Understood. When I need to be alone, I work out a few times each week." is spoken by **Q**, but it is reassigned to **R** in the corrupted version, reversing the speaker roles. Additionally, line 6 the response "A little bit. I can get into Taylor Swift.", originally served as a answer to respond the previous utterance asking: "Do you like country music?". Without this context, the dialogue loses coherence. Thus, corruption significantly disrupts conversational flow and context, negatively affecting dialogue coherence during both model training and evaluation. The evaluation using similarity to a reference utterance, described in section 3 is even more problematic when the context for the utterance has changed, so that even the reference utterance may be incoherent.

3.5. Discussion

Taken together, the four issues discussed above raise doubts about whether LAPDOG is really an improvement over a baseline system that doesn't include retrieved stories. The first step is to re-evaluate LAPDOG, both on the full, uncorrupted dialogues, and using more appropriate evaluation metrics. We describe this in the next section. We plan to address the incoherence and contraditions issues in future work, using methods inspired by cognitive and linguistic theories of relevance and coherence rather than word-similarity.

Source(model)	System	F1 ↑	BLEU ↑	ROUGE-L ↑
Ours(T5-xl)	Baseline	13.53	2.47	15.16
	LAPDOG	**16.52**	**3.56**	**15.80**
LAPDOG(T5-xl)	Baseline	16.22	3.55	15.55
	LAPDOG	**17.11**	**3.56**	**15.64**

Table 5: Automatic similarity–metric scores from our reproduction (*top*) and from the original LAPDOG paper (*bottom*) using **corrupted** dialogue; higher is better for every metric. Bold marks the better system within each source block.

4. Evaluation Methodology

To assess both the evaluation metrics and LAPDOG's reported gains, we conducted a comprehensive re-evaluation. We began by replicating the authors' experiments using their official code, confirming that our reproduction matches the published numbers and preserves the same performance trend: LAPDOG consistently surpasses the baseline (see Table 5). To extend this analysis, we downloaded the full CONVAI2 (Dinan et al., 2020) conversations via ParlAI (Miller et al., 2017) and re-trained both LAPDOG and the baseline model on this uncorrupted dataset. From this setup, we randomly selected 20 test dialogues and generated responses using the re-trained LAPDOG model, the baseline model, and the original ground-truth answers. The outputs were anonymized and shuffled before being evaluated by two human annotators—the second author and a computer science master's student—using the instructions shown in the Appendix A, Table 9. In addition, two large language models (ChatGPT-o1 and DeepSeek-R1) were used to assess response quality with prompts like the example shown in Table 6. Both human and LLM evaluations employed two measures: a 1–5 Likert scale for overall quality (with 5 as best) and a ranking scheme (1 = best, 3 = worst) to capture relative preferences even when absolute scores were similar. For comparison, we also computed the similarity metrics used in Huang et al. (2023). Finally, we computed average rating and ranking scores for both human and LLM annotations, along with an overall assessment combining all annotations.

As noted in Section 2, alignment between LLM-based evaluations and human judgments is not universal. To assess this alignment in our setting, we applied two complementary analyses. First, we computed Pearson correlation coefficients between human annotators, LLM evaluators, and similarity metrics to quantify agreement across evaluation sources. Second, we used pairwise Williams tests (Williams, 1959) to determine whether differences between correlations sharing a common variable were statistically significant. These analyses pro-

System Persona:
`persona: i love to drink fancy tea. i have a big library at home. i'm a museum tour guide. i'm partly deaf.`

Dialogue Context:
...
R: i love reading. i have a big library at my house.
Q: that's very nice, i love to read too and read about journals.
R: i've a tough time hearing so reading is my happy place.
Q: cool that is okay. my mom was very tough on me.

Candidate Responses:
R1: *"what do you do for a living?"*
R2: *"i'm sorry to hear that. i hope you have a wonderful life."*
R3: *"i work as a hearing impaired museum tour guide."*

Rating Prompt:
Please rate each candidate response (R1, R2, R3) on a scale from 1 to 5 based on overall response quality, 5 is the highest.

Ranking Prompt:
Please rank each candidate response (R1, R2, R3) from 1 to 3, where **1 is your favorite response** and **3 is your least favorite**.

Table 6: Example evaluation prompt used for LLM evaluations.

vide a principled basis for comparing the consistency of LLM and human judgments relative to traditional surface-level metrics.

5. Empirical Findings

This section reports our empirical findings on retrieval-augmented personalized dialogue from both computational and cognitive perspectives. Grounded in theories of dialogue as a collaborative process requiring coherence, consistency, and shared understanding, we assess whether current model and metrics capture these properties. We first summarize quantitative results comparing LAPDOG, the baseline, and original responses across similarity metrics and human/LLM evaluations. We then examine the alignment between human and LLM judgments and their relation to surface-level similarity metrics, revealing the latter's limitations in capturing discourse-level and cognitive qualities.

5.1. Main Results

As established in Section 4 and Table 5, our reproduction of the LAPDOG setup matches the reported results and preserves the original performance trend. Building on this, Table 7 extends the evaluation to full, uncorrupted dialogues, enabling direct comparison with the corrupted setup.

Despite overall improvement with full context, the baseline model outperforms LAPDOG on BLEU and ROUGE-L, while LAPDOG shows only a slight F1 gain. This suggests that LAPDOG's previously reported advantage may stem from dataset or context differences rather than genuine modeling improvements.

Human and LLM evaluations show a similar trend. As shown in Table 7, both groups of evaluators consistently rate the original human responses highest, followed by the baseline, and then LAPDOG. Although the baseline receives slightly better average ratings and rankings than LAPDOG, the gaps are small and not statistically significant according to one-sided Wilcoxon (Wilcoxon, 1945) signed-rank tests. The strong agreement between human and LLM evaluators indicates that they rely on similar criteria when judging response quality. These criteria may extend beyond surface similarity to include broader aspects of conversational adequacy. In the next subsection, we explore this possibility by examining how their evaluations relate to lexical similarity metrics.

5.2. Analysis of Human and LLM Judgments

To further examine the relationship between human and LLM evaluations, we analyze their pairwise correlations across all measures. Table 8 reports Pearson correlation between human ratings, LLM ratings, and automatic metrics. From a cognitive and linguistic perspective, this analysis allows us to assess whether LLMs approximate human interpretive processes–such as evaluating coherence, contextual fit, and persona consistency.

The upper part of Table 8 shows very high correlations for both ratings and rankings across human and LLM judges, with no clear distinction by evaluation categories. This indicates that human evaluators provide stable assessments of dialogue quality and that LLMs closely replicate these judgments.

Further statistical comparisons using the Williams test (Williams, 1959) (Appendix B Table 10) show no significant differences between human-LLM correlations ($p > 0.05$), confirming that their evaluations are statistically indistinguishable. Overall, these findings demonstrate that LLM-based evaluators can reliably approximate human judgments of conversational quality.

5.3. Limitations of Lexical Similarity Metrics

Lexical similarity metrics such as BLEU, ROUGE, and F1 remain widely used in dialogue evaluation,

Evaluation Results on Test Set					
Model	**Similarity Metrics**			**Overall Assessment**	
	F1	**BLEU**	**ROUGE-L**	**Rating (1-5)**	**Ranking (1-3)**
BASELINE T5$^{S}_{Sup}$(full)	17.23	**8.03**	**19.75****	**2.71**	**2.16**
T5$^{S}_{Sup}$+LAPDOG(full)	**18.06**	6.56	19.27	2.43	2.30
Original Response	–	–	–	3.96	1.41

Detailed Human and LLM Evaluation						
	Human Annotator 1		**Human Annotator 2**		**Human Avg**	
Model	**Rating**	**Ranking**	**Rating**	**Ranking**	**Rating**	**Ranking**
BASELINE T5$^{S}_{Sup}$(full)	**2.60**	2.30	**3.05**	**2.15**	**2.83**	**2.23**
T5$^{S}_{Sup}$+LAPDOG(full)	2.45	**2.25**	2.45	2.25	2.45	2.25
Original Response	3.60**	1.45**	4.10**	1.25**	3.85**	1.35**
	o1		**DeepSeek**		**LLM Avg**	
BASELINE T5$^{S}_{Sup}$(full)	**2.55**	**2.05**	**2.65**	**2.15**	**2.60**	**2.10**
T5$^{S}_{Sup}$+LAPDOG(full)	2.25	2.30	2.55	2.40	2.40	2.35
Original Response	3.95**	1.55*	4.20**	1.40**	4.08**	1.48**

Table 7: Comparison of similarity metrics versus human/LLM evaluation. Top: similarity metrics and overall assessment. Bottom: detailed breakdown of individual evaluator scores. Rating: higher is better (1-5 scale); Ranking: lower is better (1 = best, 3 = worst). **Bold** numbers indicate better performance between baseline and LAPDOG models. Asterisks denote statistical significance of the difference between three responses according to one-sided Wilcoxon signed-rank tests: $^{*}p < 0.05$, $^{**}p < 0.01$.

	LAPDOG		**Baseline**		**Original**		**Average**	
Annotator Pair	**Rating**	**Ranking**	**Rating**	**Ranking**	**Rating**	**Ranking**	**Rating**	**Ranking**
H1 & H2	0.600	0.659	0.489	0.549	0.222	0.347	0.437	0.518
H1 & DS	0.734	0.621	0.269	0.370	0.348	0.313	0.450	0.435
H1 & o1	0.580	0.459	0.348	0.315	0.330	0.096	0.419	0.290
H2 & DS	0.463	0.533	0.068	0.374	0.338	0.508	0.290	0.471
H2 & o1	0.723	0.793	0.313	0.499	0.597	0.661	0.544	0.651
DS & o1	0.791	0.575	0.143	0.440	0.570	0.515	0.501	0.510
H1 & BLEU	0.006	0.053	0.287	-0.166	-	-	0.147	-0.057
H2 & BLEU	0.104	-0.217	-0.070	0.037	-	-	0.017	-0.090
DS & BLEU	-0.053	0.037	0.086	-0.180	-	-	0.017	-0.072
o1 & BLEU	0.165	-0.253	-0.058	0.071	-	-	0.054	-0.091
H1 & ROUGE-L	0.025	-0.140	-0.043	0.563	-	-	-0.009	0.212
H2 & ROUGE-L	0.397	-0.412	-0.133	0.228	-	-	0.132	-0.092
DS & ROUGE-L	0.027	-0.092	0.139	-0.114	-	-	0.083	-0.103
o1 & ROUGE-L	0.188	-0.066	0.094	0.056	-	-	0.141	-0.005
H1 & F1	0.099	-0.124	0.088	0.479	-	-	0.094	0.178
H2 & F1	0.407	-0.372	-0.030	0.163	-	-	0.189	-0.105
DS & F1	0.044	-0.103	0.179	-0.151	-	-	0.112	-0.127
o1 & F1	0.182	-0.044	0.128	0.023	-	-	0.155	-0.011

Table 8: Pearson correlations among human annotators, LLM evaluators, and similarity metrics (BLEU, ROUGE-L, F1) across LAPDOG, baseline, and original answers. H1 and H2 refer to Human Annotators 1 and 2; DS is DeepSeek; o1 is a ChatGPT model, both LLM-based annotators. Darker shading indicates stronger correlation.

yet they provide only a limited view of conversational quality. The following analysis compares these metrics with human and LLM judgments, revealing overlap-based measures emphasize lexical resemblance while overlooking deeper qualities such as coherence, persona consistency, and pragmatic appropriateness.

The lower part of Table 8 compares human/LLM judgments with BLEU, ROUGE-L, and F1. These correlations are substantially weaker and in some cases opposite to expectation (note that we would expect negative correlations with rankings), which indicates a clear divergence between lexical overlap metrics and human or LLM evaluations. To

verify that this divergence is not due to random variation, we conducted Williams test (Appendix B Table 10) which confirms that these differences are statistically significant in nearly all cases ($p < 0.05$). Specifically, correlations involving lexical overlap metrics such as BLEU, ROUGE, and F1 differ markedly from those based on human or LLM evaluations, demonstrating that overlap-based metrics capture a fundamentally different signal from the coherence and consistency that define effective dialogue.

From a cognitive and linguistic perspective, this divergence is not unexpected. Dialogue quality depends not only on surface resemblance but on the maintenance of coherence, persona consistency, and shared understanding—properties that emerge over multiple turns rather than in local n-gram overlap. Lexical metrics, which treat language as a sequence of tokens, cannot capture how interlocutors adapt to one another's intentions, manage reference, or sustain common ground. Consequently, improvements reported under these metrics may not reflect genuine communicative competence.

Overall, the findings align with theories of dialogue that highlight coherence and common ground as key features of effective interaction. Future evaluation frameworks should therefore move beyond token-level overlap toward cognitively informed measures that capture the structural and pragmatic dimensions of natural conversation.

6. Discussion and Future Directions

In this paper, we provided a critical analysis of the LAPDOG pipeline, revealing key shortcomings in dialogue context continuity, persona consistency, and relevance. Given the issues identified in Section 3 and the re-evaluation in Section 4 and results in Section 5, we propose several directions for future improvement in RAG-based personalized dialogue using stories. First, to address the limitations discussed in Section 3.1, we recommend replacing reference-based similarity metrics (e.g., BLEU, ROUGE) with LLM-based evaluators. As discussed in section 2.3, this is generally more appropriate for evaluation of dialogue capability, and as demonstrated in section 5, there are high correlations with human judgments for evaluating persona-based dialogue.

To address the issues raised in Section 3.3, we propose improving the story selection process by incorporating a consistency classifier to filter out contradictory candidates. Contradictory information should be removed before stories are passed to the generator. We also propose categorizing candidate external sources based on their relationship to the persona: overlap, complementarity, and independence. Overlap refers to stories that restate persona facts and are already reflected in the profile—these offer minimal added value. Complementary stories enrich the persona with related but non-redundant information, helping expand depth without contradiction. Independent stories introduce new, unrelated topics and may add conversational breadth. We aim to study the impact of each category on coherence and personalization.

To mitigate the coherence problems discussed in Section 3.2, we recommend incorporating metrics that analyze relevance and fluency, such as the FED fluency score FED (Mehri and Eskenazi, 2020). We have addressed the issue in Section 3.4 by using the full dialogue context from the original CONVAI2 dataset, and re-evaluated in Section 4. Together, these contributions chart a clear path toward more coherent, persona-aware RAG dialogue systems that better integrate external story knowledge and align with human expectations.

Finally, it will be important to test the models in actual dialogue with the intended user population, rather than just offline analysis of individual dialogue turns.

7. Conclusion

This work re-evaluated retrieval-augmented personalized dialogue through a cognitive and linguistic lens, using LAPDOG as a representative case. We introduced a systematic evaluation framework that combines human and LLM judges with analysis methods grounded in theories of dialogue, emphasizing coherence, persona consistency, and engagement. Empirically, we found that human and LLM judgments align closely but diverge sharply from lexical similarity metrics, revealing a gap between linguistic form and communicative function. Building on these insights, we point to practical directions for system design. This moves retrieval-augmented dialogue systems toward behaviors that better reflect principles of natural human communication.

8. Limitations

Our work has several limitations. First, our hybrid human+LLM evaluation was tested using only two LLMs and two human raters on 20 example situations taken from the CONVAI dataset and augmented with stories from ROC-Story. It is unclear how well the results generalize to use of different resources. Second, our proposals still need to be fully implemented and validated to show improvements over LAPDOG and the baseline persona usage. Future work should therefore expand the evaluation and run additional experiments to confirm the reliability and usefulness of the proposed framework.

9. Acknowledgements

This work was supported by the U.S. Army Research Office under Cooperative Agreement Numbers W911NF-20-2- 0053 and W911NF-25-2-0040.

Bibliographical References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Xueliang Cheng, Ziwei Ji, Jiashuo Wang, Zheng Gong, Yuxuan Lai, and Baobao Chang. 2024. In-dialogue learning for personalized dialogue generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10448–10465, Miami, Florida, USA. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC.

Simon Garrod and Martin J. Pickering. 2009. Joint action, interactive alignment, and dialogue. *Topics in Cognitive Science*, 1(2):292–304.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540, Singapore. Association for Computational Linguistics.

Zhihao Li, Yixin Zhang, Qian Chen, and Wenqiang Lei. 2024. Unims-rag: Unifying multi-source retrieval-augmented generation for personalized dialogue. *arXiv preprint arXiv:2401.13256*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016a. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016b. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1116–1126.

Fanghua Lu, Yizhe Zhang, Xiang Li, and Minlie Huang. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. *arXiv preprint arXiv:2310.18342*.

Bhaskar Komaragiri Majumder, Y-Lan Boureau, Jason Weston, and Antoine Bordes. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers*, pages 604–611.

Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. 2025. A comparative analysis of metaphorical cognition in ChatGPT and human minds. *Cognitive Computation*, 17(35):1–12.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Jihyun Oh, Chan-Hyeong Park, Sang-Hyun Park, and Dongchan Kim. 2023. Pk-icr: Persona-knowledge interactive context retrieval for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Massimo Poesio and Hannes Rieser. 2001. Completions, coordination, and alignment in dialogue. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Aalborg, Denmark. Association for Computational Linguistics.

Matthias Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. arXiv preprint arXiv:2305.15780.

Shikhar Sharma, Hannes Schulz, Maxine Eskenazi, and Yoshua Bengio. 2017. Natural language understanding for task-oriented dialogues in the presence of out-of-domain utterances. In *Proceedings of the 8th International Workshop on Spoken Dialogue Systems (IWSDS)*.

Chrysanthi Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Rethinking the evaluation of dialogue systems: Effects of user feedback on crowdworkers and llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142.

Chen Tang, Chengguang Tang, Yating Zhang, Jinchao Zhang, and Jie Zhou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv preprint arXiv:2305.11482*.

Yi Wang, Qian Liu, Yunchang Cui, Yining Wang, and Tatsunori Hashimoto. 2023. Safari: Large language models as source planner for personalized knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4129–4143.

Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, Mark Cusick, Param Kulkarni, Zhengping Ji, Yasser Ibrahim, and Xia Hu. 2025. Dhp benchmark: Are llms good nlg evaluators?

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

E. J. Williams. 1959. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399.

Yunhe Xie and Rui Mao. 2025. PGIF: A personality-guided iterative feedback graph network for multimodal conversational emotion recognition. *IEEE Transactions on Computational Social Systems*, pages 1–13.

Yuxin Ye, Zhaoqing Li, and Guodong Zhou. 2023. Retrieval-enhanced generation for personalized dialogue: Balancing consistency and diversity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pages 9675–9690, Toronto, Canada. Association for Computational Linguistics.

Xiaoyu Zhan, Wenhao Chen, and Qiang Liu. 2024. Contextual rag: Enhancing coherence in retrieval-augmented dialogue generation. *IEEE Transactions on Affective Computing*.

Luyao Zhu, Wei Li, Rui Mao, and Erik Cambria. 2024a. HIPPL: Hierarchical intent-inferring pointer network with pseudo labeling for consistent persona-driven dialogue generation. *IEEE Computational Intelligence Magazine*, 19(4):63–78.

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea, and Erik Cambria. 2023. PAED: Zero-shot persona attribute extraction in dialogues. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), volume 1, page 9771–9787, Toronto, Canada. Association for Computational Linguistics.

Luyao Zhu, Rui Mao, Erik Cambria, and Bernard J. Jansen. 2024b. Neurosymbolic ai for personalized sentiment analysis. In *HCI International 2024 – Late Breaking Papers*, volume 15119 of *Lecture Notes in Computer Science*, pages 269–290, Washington, DC, USA. Springer.

Language Resource References

Dinan, Emily and Logacheva, Varvara and Malykh, Valentin and Miller, Alexander and Shuster, Kurt and Urbanek, Jack and Kiela, Douwe and Szlam, Arthur and Serban, Iulian and Lowe, Ryan and Prabhumoye, Shrimai and Black, Alan W. and Rudnicky, Alexander and Williams, Jason and Pineau, Joelle and Burtsev, Mikhail. 2020. *ConvAI2: The Second Conversational Intelligence Challenge Dataset*. Springer. Dataset accompanying the NeurIPS ConvAI2 competition, based on Persona-Chat dialogues.

Miller, Alexander H. and Feng, Will and Fisch, Adam and Lu, Jiasen and Batra, Dhruv and Bordes, Antoine and Parikh, Devi and Weston, Jason. 2017. *ParlAI: A Dialog Research Software Platform*. Association for Computational Linguistics. Available at `https://parl.ai`.

Appendices

A Evaluation Instructions

Table 9 gives the instructions and one dialogue example for human rating and rankings.

Section	Content / Evaluation Fields
Task Definition	For every test dialogue you will be shown the system's **persona**, the recent **dialogue context**, and three **candidate responses** (`R1`, `R2`, `R3`). Annotate each response based on: 1. **Absolute Quality (1–5):** **5 Excellent:** Coherent, engaging, on-topic, persona-consistent. **4 Good:** Minor wording or relevance issues. **3 Adequate:** Bland or one noticeable flaw. **2 Poor:** Incoherent or ignores persona/context. **1 Unusable:** Nonsense, contradictory, or toxic. 2. **Relative Rank (1, 2, 3):** 1=best, 2=middle, 3=worst. Ties are allowed.
System Persona	`i love to drink fancy tea. i have a big library at home. i'm a museum tour guide. i'm partly deaf.`
Dialogue Context	**Q:** hi how are you doing? i am okay how about you? **R:** i am doing good. just sipping tea. what do you do for work? **Q:** i used to do home health aide but now i am disabled. **R:** i am sorry to hear that. what happened **Q:** i graduated and got my license, worked a while, and became ill. **R:** so what do you do now for fun? i like to read. **Q:** i have dogs and i walk them. and a cat. **R:** what kind of dogs do you have? **Q:** i enjoy reading—read about michael jackson and psychology today. **R:** i love reading. i have a big library at my house. **Q:** that's very nice, i love to read too and read about journals. **R:** i've a tough time hearing so reading is my happy place. **Q:** cool that is okay. my mom was very tough on me.
Response 1 (R1)	*"what do you do for a living?"* **Quality (1–5):** _____ **Rank (1–3):** _____
Response 2 (R2)	*"i'm sorry to hear that. i hope you have a wonderful life."* **Quality (1–5):** _____ **Rank (1–3):** _____
Response 3 (R3)	*"i work as a hearing impaired museum tour guide."* **Quality (1–5):** _____ **Rank (1–3):** _____
Quality Legend	**5:** Excellent (coherent, engaging, persona-consistent) **4:** Good (minor wording/relevance issues) **3:** Adequate (bland or one noticeable flaw) **2:** Poor (incoherent; ignores persona/context) **1:** Unusable (nonsense, contradictory, or toxic)

Table 9: Human Evaluation Sheet

B Correlation Difference Testing

Table 10 shows Williams tests results for correlations between variables, showing that human and LLM correlations are significantly higher than correlations with similarity metrics.

Shared variable: H1

Comparison	r_1	r_2	t	p
H1–o1 vs. H1–BLEU	0.487	0.081	2.12	0.041*
H1–DS vs. H1–BLEU	0.571	0.081	2.64	0.012*
H1–H2 vs. H1–BLEU	0.555	0.081	2.48	0.018*
H1–o1 vs. H1–ROUGE	0.487	0.038	2.33	0.026*
H1–DS vs. H1–ROUGE	0.571	0.038	2.94	0.006*
H1–H2 vs. H1–ROUGE	0.555	0.038	2.91	0.006*
H1–o1 vs. H1–F1	0.487	0.073	2.15	0.038*
H1–DS vs. H1–F1	0.571	0.073	2.81	0.008*
H1–H2 vs. H1–F1	0.555	0.073	2.74	0.009*
H1–H2 vs. H1–o1	0.555	0.487	0.67	0.510
H1–H2 vs. H1–DS	0.555	0.571	-0.13	0.899

Shared variable: H2

Comparison	r_1	r_2	t	p
H2–DS vs. H2–BLEU	0.309	0.015	1.37	0.179
H2–o1 vs. H2–BLEU	0.552	0.015	2.93	0.006*
H2–DS vs. H2–ROUGE	0.309	0.149	0.77	0.447
H2–o1 vs. H2–ROUGE	0.552	0.149	2.21	0.034*
H2–DS vs. H2–F1	0.309	0.165	0.71	0.481
H2–o1 vs. H2–F1	0.552	0.165	2.13	0.040*
H2–H1 vs. H2–o1	0.555	0.552	0.03	0.974
H2–H1 vs. H2–DS	0.555	0.309	2.18	0.035*
H2–DS vs. H2–o1	0.309	0.552	-2.00	0.053

Shared variable: o1

Comparison	r_1	r_2	t	p
o1–DS vs. o1–BLEU	0.523	0.102	2.19	0.035*
o1–DS vs. o1–ROUGE	0.523	0.092	2.30	0.027*
o1–DS vs. o1–F1	0.523	0.091	2.35	0.024*
o1–DS vs. o1–H1	0.523	0.487	0.34	0.733
o1–DS vs. o1–H2	0.523	0.552	-0.22	0.826

Shared variable: DS

Comparison	r_1	r_2	t	p
DS–o1 vs. DS–BLEU	0.523	0.050	2.52	0.016*
DS–o1 vs. DS–ROUGE	0.523	0.095	2.28	0.029*
DS–o1 vs. DS–F1	0.523	0.133	2.08	0.044*
DS–o1 vs. DS–H1	0.523	0.571	-0.45	0.652
DS–o1 vs. DS–H2	0.523	0.309	1.77	0.084

Table 10: Williams tests for differences between two correlations that share one variable (40 paired dialogues). An asterisk (*) indicates significance at $\alpha = 0.05$ (two-tailed).